Review of Montana Department of Environmental Quality 2012 "Scientific and Technical Basis of the Numeric Nutrient Criteria for Montana's Wadeable Streams and Rivers: Addendum 1"

General comments:

In preparation for the review below, I read Suplee et al. 2012 "Scientific and Technical Basis of the Numeric Nutrient Criteria for Montana's Wadeable Streams and Rivers: Addendum 1", Suplee et al. 2011 "Assessment Methodology for Determining Wadeable Stream Impairment Due to Excess Nitrogen and Phosphorus Levels", and reviewed Suplee et al. 2005 " Identification and Assessment of Montana Reference Streams: A Follow-up and Expansion of the 1992 Benchmark Biology Study." In addition, I reviewed considerable literature to refresh my memory about details and look for additional information.

Overall, the "Scientific and Technical Basis of the Numeric Nutrient Criteria for Montana's Wadeable Streams and Rivers: Addendum 1" was scholarly, thorough, and as scientifically sound as any state nutrient criteria document that I've reviewed. I think the approach is sound, information is relatively sufficient, and this work provides a very good next step for stakeholders of Montana and the development of their nutrient criteria to protect their resources. I have great respect for the originality and scientific rigor of the research conducted by MDEQ and its application in water policy. With that said, my responsibility is to indicate strengths and weaknesses in the approach and results, as well as address a set of specific questions. My review includes recommendations for additional approaches and sources of information for deriving benchmarks for nutrient criteria and selecting benchmarks for criteria that I hope MDEQ will find useful in revisions of this document or in their future work. Much research remains to refine the information needed for states and tribes to establish nutrient criteria that will adequately protect designated uses of their waters without overprotection. MDEQ is a leader in that effort and that effort serves the state of Montana well.

Below I've addressed the specific review questions and commented on related issues.

MT – Wadeable Streams Draft Peer Review Questions

1. **MDEQ is considering two approaches for the derivation of numeric nutrient criteria in wadeable streams: (1) eco-regional reference condition, and (2) regional and non-regional stressor-response studies. Compare and contrast the ability of each approach to provide a sound scientific basis for numeric nutrient criteria derivation. Please provide documentation on any identified ranges protective of aquatic life based on**

similar studies. If possible, please provide alternate methodologies using available data and tools, and describe the corresponding advantages and disadvantages.

MDEQ used ecoregion specific stressor-response relationships and ecoregional reference condition to derive numeric nutrient criteria for wadeable streams. Stressor-response relationships were used to determine the nutrient concentration at which undesirable effects in stream condition occurred. Ecoregional reference condition was used to determine the range of nutrient conditions at groups of sites with minimally impacted condition (*sensu* MDEQ 2005), that meet designated uses, and that have similar natural determinants of ecological condition (based on ecoregion constraints). Combining information from stressor-response relationships and ecoregional reference condition, nutrient criteria were then proposed for nutrient concentrations (both TP and TN) that were: 1) related to negative effects in biological condition that were predicted by stressor-response relationships and 2) greater than or equal to the 75th percentile of nutrient concentrations observed at reference sites. If sufficient knowledge is available for characterizing responses of valued ecological attributes (e.g. biological condition) to nutrient enrichment and minimally impacted nutrient concentrations at reference sites, and these characterizations are done appropriately, then I would argue that this is the best framework for deriving nutrient criteria. So an appropriate question to ask is, "Has MDEQ appropriately characterized nutrient concentrations in minimally impacted condition and responses of valued ecological attributes (e.g. biological condition and other indicators of designated use support) to nutrient enrichment?" I'll get to that question later after I briefly defend the MDEQ approach.

I have argued that nutrient criteria (and other stressor criteria) for a site should be derived with at least three steps (e.g. Stevenson et al. 2004, 2008; Soranno et al. 2008), given sufficient information:

1. determine expected conditions[1] for a site (which can be reference or desired conditions) based on management goals (which can be designated uses);
2. determine effect of nutrient concentrations on valued ecological attributes related to management goals for the site (e.g. biological condition or other indicators of designated uses) and select benchmarks in nutrient concentrations for possible criteria;
3. select benchmarks in nutrient concentrations that are greater than or equal to minimally disturbed condition and at concentrations with acceptable risk to impairment of valued ecological attributes (i.e. often measures or indicators of designated uses).

---

[1] Expected condition can be defined as minimally disturbed, least disturbed, best available, or desired condition (Stevenson et al. 2004). Here I use definitions of, least disturbed, best available from Stoddard et al. (2006) such that: minimally disturbed is "the condition of streams in the absence of significant human disturbance;" least disturbed is "found in conjunction with the best available physical, chemical, and biological habitat conditions given today's state of the landscape;" and best attainable condition is *"equivalent to the expected ecological condition of least-disturbed sites if the best possible management practices were in use for some period of time."* Desired condition is related to natural resources management and specifically addresses situations in which we management for attributes that may not be greatest in minimally disturbed conditions

In step 1 we should characterize the reference or desired (=expected) condition for the site that should include all physical, chemical and biological conditions that are related directly or indirectly to our management goals (e.g. designated uses) and that occur within the water, the riparian zone, the watershed, regionally, and even globally for contaminants transported through the atmosphere from distant sources. In some special cases, our goals may be to manage for desired condition (*sensu* Stevenson et al. 2004), such as more productive fisheries that are not characteristic of minimally disturbed conditions with high levels of biological condition (*sensu* Davies and Jackson 2006) in naturally low productivity ecosystems. Thus, tradeoffs between managing for productive fisheries and high levels of biological condition (biological integrity) are likely and should be addressed with tiered uses and different tiered uses for different waters within a region that meet the needs of regional stakeholders (Stevenson and Sabater 2010). Also, natural variation in climate, geology, hydrology, and water chemistry cause variation in minimally disturbed condition among ecoregions and among sites (e.g. Cao et al. 2007; Hawkins et al. 2010). So expected condition and nutrient critera, eventually, should be derived separately by ecoregion or by sites (e.g. Herlihy and Sifneos 2008; Soranno et al. 2008; Suplee et al. 2012 (the MDEQ document being reviewed)).

In step 2, we determine relationships between valued ecological attributes indicating designated and desired use support and nutrient concentrations. Nutrient concentrations are not a valued attribute because most people do not value them directly and only perceive risk from them if they cause problems to ecosystem services they do care about. There is little public support for managing nutrients independently of the effects that nutrients have on valued ecological attributes. We should not use reference condition nutrient concentrations alone to derive nutrient criteria because: 1) without stressor-response relationships we cannot be sure that nutrients affect valued attributes of the ecosystem and 2) we don't know the effects of incrementally increasing nutrient concentrations and at what nutrient concentrations risk of losing attributes become unacceptable. In evaluating stressor-response relationships, nutrient concentration benchmarks for potential criteria should be identified at the highest levels of nutrient concentrations at which an acceptable risk of losing valued attributes occurs. Thresholds in stressor-response relationships are highly valuable for delineating levels of nutrient concentrations at which risk levels change dramatically, thereby generating consensus among stakeholders for establishing criteria at specific nutrient concentration benchmarks.

In step 3, we determine what responses in valued ecological attributes change have acceptable risk benchmarks at nutrient concentrations greater than or equal to expected (usually reference) and then determine which benchmarks should be selected for nutrient criteria. In general, it's impractical (although not impossible) to manage a resource for nutrient concentrations lower than minimally or least disturbed condition, so nutrient criteria are usually at least as high as nutrient concentrations in reference conditions[2]; and criteria may be higher than reference conditions if valued attributes are not affected by nutrient concentrations less than or equal to reference conditions.

[2] Nutrient concentrations characteristic of reference conditions and supporting conditions of reference conditions are not any concentration within the range of nutrient concentrations observed at reference conditions. This will be discussed later in the text.

Now to the question, "Has MDEQ appropriately characterized nutrient concentrations in minimally impacted condition and responses of valued ecological attributes (e.g. biological condition and other indicators of designated use support) to nutrient enrichment?" Here I will also address elements of the review question:

- Please provide documentation on any identified ranges protective of aquatic life based on similar studies.
- If possible, please provide alternate methodologies using available data and tools, and describe the corresponding advantages and disadvantages.

I'll address the question and review question elements by criteria development step, and change the order of steps to correspond to the MDEQ methodology (characterizing stressor-response relationships and reference condition, and then deriving criteria).

Characterizing stressor-response relationships. MDEQ relies heavily on the relationships between nutrient concentrations and chlorophyll a, chlorophyll a and DO stress, and chlorophyll and aesthetics to related nutrient concentrations to support of designed uses. The nutrient-chlorophyll relationship is therefore the primary determinant of DO stress (e.g. Stevenson et al. 2012), which is an important stress on aquatic biota. The nutrient-chlorophyll relationship is also a primary determinant of aesthetics issues. Suplee et al. (2009) show reduced desirability of rivers for recreations use with chlorophyll a exceeding 125-150 mg chl a m^{-2}. The stressor-response relationships that they use are peer-reviewed and scientifically sound, or they have been developed by their own research in regions in which they have particular concern that that existing nutrient-response relationships would not apply. They consider different stressor-response relationships for different ecoregions, which is appropriate, because we would not expect high gradient streams, as in the mountains or foothills, to respond the same to nutrient pollution as in the low gradient streams of the prairies (see Stevenson et al. 2006 for example or ecoregion specific relationships). As an aside, I tried to compare the nutrient concentrations required to produce 125 mg chl a m^{-2}, but I could not determine which equation in Dodds et al. 2006 was equation 19. Comparing predicted nutrient concentrations at chlorophyll management targets using models in Mebane et al. (2009), Dodds et al. (1997 and 2006), and Stevenson et al. (2006) would be informative. Providing these models in the report would have been valuable for establishing the basis for the range in nutrient concentrations that were reported as required to maintain 125 mg chl a m^{-2}. Also, although results of experiments are based on soluble nutrients, Bothwell's experimental work with P and the N and P experimental work of Rier and Stevenson (2006) could be used to support determination of nutrient benchmarks for regulating chlorophyll a accrual.

While MDEQ's approach is scientifically sound, there are other relationships between nutrients and elements of stream ecosystems that may be important for determining whether nutrient pollution threatens designated uses of Montana waters. MDEQ definition of minimally impacted condition[3] indicates that more than chlorophyll and DO stress on invertebrates

[3] MDEQ (2005, p 2) defines minimally impacted condition as "Tier 2 — Minimally Impacted Condition" as "The characteristics of a waterbody in which the activities of man have made small changes that do not affect the completeness of the biotic community structure and function and the associated physical, chemical, and habitat conditions, and all numeric water quality standards are met and all beneficial uses are fully supported unless

should be included in stressor-response relationships. Since I did not find reference to the attributes specifically used to characterize designated uses of MT waters, I will mention some additional information that may be valuable to consider and which might not have been considered by MDEQ.

Relationships between nutrient concentrations, chlorophyll a, DO stress, and aesthetics likely cover most designated uses related to recreation, but may not protect biological condition of invertebrates, algae, and ecosystem function. Stevenson et al. (2008) observed very sensitive response of benthic diatom assemblages in the high gradient streams of the mid-Atlantic highlands with loss of sensitive species and deviations in species composition from reference condition at nutrient concentrations well below the 30 µg TP/L benchmark used for several MT ecoregions. With the abundance of periphyton data in the Western EMAP, the STAR reference site projects (Hawkins et al.), and now the National Rivers and Streams Assessment, generating informative stressor-response relationships for biological condition of periphyton and nutrients should be very practical.

In addition, Miltner and Rankin (1998), Yuan (2004), Smith et al. (2007), and Wang et al. (2007) describe invertebrate responses to nutrient concentrations that could be used to justify benchmarks for protecting biological condition of invertebrate communities. The mechanisms causing changes in species composition at relatively low nutrient concentrations are not well understood. DO and pH stress with nutrient enrichment are two likely mechanisms (Stevenson et al. 2012). In addition, release of streams and rivers from nutrient limitation enables invasion of habitats by taxa requiring higher productivity levels to survive and may shift competitive hierarchies in ways that cause loss of sensitive taxa adapted to naturally stressful low nutrient concentrations (Stevenson et al. 2008). Finally, release of aquatic ecosystems from nutrient limitation may enable invasion and reproduction of aquatic bacteria and fungi that could stress all other biota.

I applaude MDEQ's use of both TN and TP criteria because either can be limiting algal growth in streams with different geological conditions and resulting water chemistry, and at different times of years in some watersheds. I think this is largely done correctly given the amount of information available, where in high P reference regions MDEQ proposes low N criteria to constrain algal accrual. I think selected concentrations will be protective of high biomass in most cases where low N is used to constrain algal accrual. However, I do want to caution that we need to learn more to accurately quantify algal nutrient relationships with both TN and TP in the model, as was used by Dodds et al. (2002, 2006) and MDEQ. Such models violate Leibig's Law of the Minimum. MDEQ does address this in their report, but in reality, those justifications may not be sufficient. There is evidence in recent research that Leibig's Law of the Minimum does not hold, which makes me think algal biomass models with TN and TP linked are appropriate. Even though the science is a bit soft here, I would recommend using

---

measured impacts are clearly linked to a natural source. Minimally impacted conditions can be used to describe attainable biological, chemical, physical, and riparian habitat conditions for waterbodies with similar watershed characteristics within similar geographic regions and represent the water body's best potential condition."

the linked models and unlinked models as a multimodel approach for getting a range of conditions that would probably constrain algal biomass below the 125-165 μg chl a m^{-2} targets.

Characterizing reference condition. MDEQ's characterization of nutrient concentrations at reference sites suffered from a low sample size in three ways: 1) for all but a couple ecoregions, there were very small numbers of sites; 2) ror a couple ecoregions, there were fewer than 30 observations of nutrient concentrations at reference sites; and 3) repeated measures of nutrient concentrations at the same site are not independent. In the truest sense of pseudoreplication, the characterization of central tendency and variation in nutrient concentrations at reference sites suffers from some level of dependence in the samples.

The pseudoreplication issue should be addressed in a straightforward manner and put into a broader context so that it does not become overly important as a distraction from the relatively sound science that does underpin MDEQ's efforts. Although Suplee et al. (2011) address the pseudoreplication issue in another report, the key point is that it should be addressed. The broader context should include the following points. First, precise characterizations of percentiles are not that important because reference condition was used as a point of "reference" for nutrient benchmarks in stressor-response relationships where undesireable conditions developed. Second, the relative independence of repeated measures in reference condition is probably pretty low, given other sources of variability in estimates of nutrient concentrations in a stream: spatial and temporal variability in nutrient concentrations of streams and analytical error. I have argued this myself (Stevenson et al. 2006). However, repeated measures statistics can be calculated relatively easily to determine the relative dependence of measurements from the same site given overall variability and to correct estimates of variance among sites for dependency in repeated measures to more accurately characterize the central tendency and variation in nutrient concentrations at reference sites. The evenness approach (calculating evenness of measures among sites) that MDEQ uses is interesting, but it does not address pseudoreplication and dependent measurement issue directly.

Modeling expected nutrient concentrations at sites with land use-nutrient relationships is another method for characterizing the central tendency and variation in nutrient concentrations in minimally disturbed conditions. Modeling reference condition is valuable when the number of reference sites is low or quality of references sites varies between regions, which may have been the case in MT. Examples of different approaches for this kind of modeling can be found in Dodds and Oakes (2004), Herlihy and Sifneos (2008), Stevenson et al. (2008), and Soranno et al. (2008).

Typically, if an endpoint of management is used in criteria development, or as pseudocriteria, as chlorophyll a, then reference condition of that parameter is also described. Reference conditions were reported consistently for TP and TN concentrations. I'd recommend that chlorophyll a, diatom decreasers metric, and Hilsenhoff's biotic index (HBI) be described for reference conditions.

Selecting nutrient benchmarks for criteria. In general, if valued ecological attributes (direct indicators of designated use support) respond sensitively within the range of nutrient conditions at reference conditions, it is difficult to justify higher nutrient benchmarks than the

75^{th} percentile of reference condition, assuming reference condition supports designated uses as described by MDEQ. I remember three distinct exceptions to this rule in MDEQ's proposed criteria. One is several Rockies ecoregions, in which proposed nutrient criteria were substantially above background concentrations, the other was in an ecoregion in which P was high and the TN criterion was well above the 75^{th} percentile of reference condition, and the other was in the River Breaks region in which no criteria were proposed because no known ecological responses to nutrients were known for concentrations that high. I'll address the River Breaks situation below with the specific question asked for the review.

I'm concerned about selecting nutrient criteria above background concentrations in the Rockies ecoregions because proposed criteria would not protect sensitive, low nutrient diatom taxa, ecosystem functions of low productivity systems, and likely corresponding biodiversity of other groups whose response to low nutrient concentrations are poorly understood (bacteria, meiofauna, even benthic macroinvertebrates species). In Stevenson et al. (2008) we observed substantial changes in species composition of diatom assemblages at low nutrient concentrations and substantial loss of sensitive, low nutrient taxa (from counts) across the range of nutrient conditions. I've seen the loss of sensitive, low nutrient taxa (from counts) with low levels of nutrient enrichment in the extensive ecological assessment work that I've done around the country. Yes, this is just loss of taxa from counts, and we're not quite sure what that means (although my students and I are trying to understand that more), but we may actually be losing more taxa from the habitat (not just counts), as well as losing fewer. At this point, it just depends upon the weight of assumptions in the model. But if this is true for diatoms, then what about other groups. Also allowing higher N concentrations as well as P concentrations could impair biological integrity of these minimally disturbed, near-natural systems. For example, releasing N limited systems from severe N limitation could cause loss of diatoms with N-fixing cyanobacterial endosymbionts (e.g. *Epithemia*) or allow invasion of potentially nuisance taxa. Also, the relaxed P and N criteria are close to thresholds for releasing systems from severe nutrient constraint, so nuisance growths of algae could occur more frequently than if nutrient criteria were constrained to reference condition. Quantifying acceptable risk of nuisance growths should guide considerations.

Setting stressor criteria at a stressor level predicted to cause a target responses (i.e. nutrient criteria at nutrient concentrations at which a model predicts a target 125 mg chl a m^{-2}) means that when the stressor is at that level, the response with be greater than the target 50% of the time and less than the target by 50% of the time and by a magnitude that is related to the mean square error of the predicted values. Should quantile regression or conditional probabilities (Paul and McDonald 2005) be used to determine the stressor level that will manage the response with an acceptable frequency and intensity of exceedance?

Thresholds, relatively abrupt changes in rates of response along stressor gradients, are valuable for deriving environmental criteria (Muradian 2001). They identify benchmarks for possible nutrient criteria and help determine which benchmarks should be used as criteria. Some threshold responses are more valuable than others (Stevenson et al. 2008). Information for different threshold responses should be interpreted differently. For example, a response showing assimilative capacity and then a threshold response as stressors increase (Stevenson et al. 2008, Figures 2A & B, , sometimes called a type III response (http://en.wikipedia.org/wiki/Functional_response)) is particularly valuable for deriving criteria because stressor levels just below the threshold are clearly protective of reference condition and provide a margin of safety. Thresholds in responses showing high rates of change at low stressor levels and little change at high stressor levels, sometimes called a saturation curve or type II response (Stevenson et al. 2008, Figure 2C, (http://en.wikipedia.org/wiki/Functional_response), is more difficult to apply in criteria development. Nutrient uptake and algal growth have this type response along nutrient gradients with highly sensitive responses to nutrients at low concentrations and little response to nutrients at high concentrations.

The question then becomes, "How low should you set criteria to constrain growth in a part of the curve that is relatively linear?" Setting criteria just below the threshold (or breakpoint as described in Dodds et al. 2002, 2006 and as applied in this MDEQ document) provides little protection from adverse effects. Algal growth and accrual are largely at their greatest levels at nutrient concentrations just below those breakpoints. So in Transitional Level IV Ecoregions of the Northwestern Glaciated Plains, a justification for TN criteria of 560 μg/L was 560 was lower than the maximum saturation threshold (with saturation thresholds of 367 and 602), this criterion would not constrain algal accrual if these models are correct (which is the assumption of using them). The real explanation for choosing that level would seem to be that 560 μg TN/L is close to the 75th percentile of reference condition and you can't expect to do much better than that, even though biomass-nutrient models indicate biomass accrual could be near maximum levels at that TN concentration.

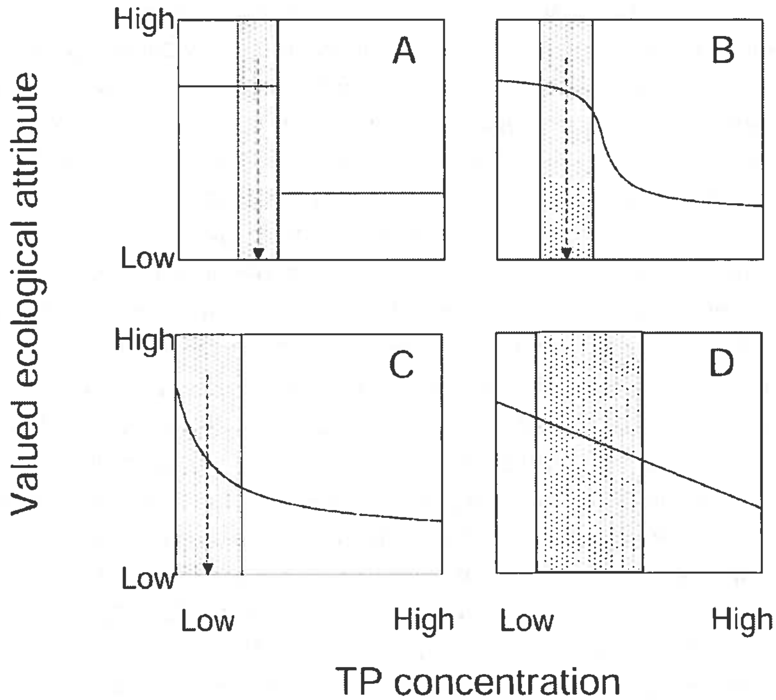

FIG. 2. Approaches to development of stressor criteria when potential responses of valued ecological attributes to stressors (e.g., total P [TP]) are nonlinear with assimilative capacity for increases at low levels the stressor (A, B), nonlinear with strong sensitivity to changes at low levels of the stressor (C), and linear (D). A stressor criterion is established at a level of a stressor that protects the valued ecological attribute. Arrows indicate TP criteria justified on the basis of the form of the stressor–response relationship. Shaded areas indicate the range of TP criteria that could be acceptable. Acceptable ranges vary as a function of the linearity of the stressor–response relationship and the type of nonlinear relationship.

Implementation. I like the concept of Tier I and II assessments for determining whether sites meet nutrient criteria. This does relate to an issue about risk of use support and the way we use statistics to define reference condition and determine whether a site meets its water quality criteria. I'll discuss this under a later question about exceedance frequencies. In particular, I like the use of biological condition assessment in the Tier II assessments with the diatom and macroinvertebrate metrics. These metrics should provide a temporally integrated signal that should complement the temporally variable assessments of nutrient concentrations. I would, however, recommend that MDEQ include metrics that evaluate decrease in sensitive native taxa as well as the increase nutrient pollution tolerant taxa (e.g. diatom increases and HBI taxa), because these taxa are key elements of biological condition for which we manage waters (Davies and Jackson 2006).

MDEQ chose to implement criteria during the growing season only, which assumes that the mechanism by which nutrients affect designated uses is by stimulating algal growth and that recreational use exposure is during the growing season. This is likely true, that algae do not bloom to nuisance levels or threaten low DO or high pH during non-growing seasons. But there are other potential ways that nutrient pollution can affect aquatic life use, which are poorly understood and poorly documented (i.e. shifts in competitive hierarchies and disease), and some nuisance growths of diatoms that alter habitat structure can occur during cooler seasons of the year.

I'm surprised there is little difference in when the growing season occurs. Why not use water temperature (for algal endpoints) and degree days (for invertebrate endpoints)? Do these time periods allow for interannual variation?

Tiered aquatic life uses should be considered (Davies and Jackson 2006). The problem with potential management challenges in the Rockies (as well as elsewhere), where reference nutrient conditions seem really low and well below most targets for designated use, plus the desire for P enrichment to support fisheries and limit *Didymosphaenia* blooms, is that we could lose attributes of natural systems that now exist. Tiered aquatic life use policies could allow protection of some systems within those ecoregions for near natural structure and function (which now exist) and allow other systems to be managed for fisheries and *Didymo* control.

I sense several issues touch on the policy doctrine of "independent applicability" of stressor and response criteria. Conceptually, one reason to set criteria within reference conditions (with a margin of safety) is because there may be negative responses that we don't know about if stressors are higher. In a perfect world, we know all the possible responses to stressors, so we could relax stressor criteria to levels that protect desired responses with acceptable risk. BUT, do we know enough about nutrient effects on designated uses to make relax criteria to ranges outside the reference condition (i.e. greater than the 75th percentile of reference condition as argued below)? MDEQ does use elements of independent applicability in their assessments. For example, level I assessments only involve comparisons of nutrients and not biological endpoints to nutrient criteria. In addition, if either N or P fail, then the system is not in compliance (MDEQ 2011, pp. 3 and 4), both do not need to fail to be in noncompliance. I did not find and review the assessment methodologies for level II decisions with sufficient detail to evaluate issues related to independent applicability.

MDEQ is as knowledgeable about potential nutrient impacts as any other state or tribal agency. They have chosen a level of risk with which they are comfortable for protecting their waters. Rivers are relatively resistant ecosystems. If errors are made, given the MDEQ "good-faith" effort, designated uses of the rivers should be able to be restored, unless there is regional extirpation of taxa which is unlikely in the short term.

2. **In Section 3.6.1., Montana suggests that no nutrient criteria are needed for streams in the 1 Level IV Ecoregion within the Northwestern Great Plains: River Breaks (43c). The MDEQ rationale for this decision is: "This level IV ecoregion has highly turbid, flashy streams with naturally elevated TP and TN levels. Concentrations observed in the region's reference sites indicate that nutrient concentrations here are already naturally elevated above the harm-to-use thresholds identified for the plains region as a whole. As such, no nutrient criteria are recommended for streams within this level IV ecoregion." Please comment on whether the state has provided a sufficient scientific basis that 1) these levels are <u>naturally</u> elevated, 2) additional increase in nutrients would not cause harm to aquatic life, and 3) that, therefore, criteria are not needed. Is the reviewer aware of any additional information that could be provided to either support the State's assessment of natural background or that could be used to derive site specific criteria?**

I don't like the idea that there are no nutrient criteria set for waters, even given the rationale that natural concentrations are naturally high and no instream or downstream effects are expected to occur. It makes me nervous that we know enough about nutrient-stream relationships to make that call. Will antidegradation policy prevent this system from getting worse? Why not have criteria be existing condition, i.e. the reference condition, as the criterion? Independent applicability would call for using reference condition of a contaminant in this case. Addressing questions 1-3. 1) I am not convinced that MDEQ has provided a sufficient scientific basis that nutrient concentrations are naturally elevated because: the quality of reference sites relative to land use in this region is not described, so how minimally disturbed is the reference condition; how is minimally disturbed and meeting designated used defined in this ecoregion if the systems are so naturally stressed; the number of reference streams sampled is low (n=8), even though the number of samples in relatively high (n=29), but note the 3 outlying samples with TP > 3.6^{10} (i.e. >3981 µg TP/L) that are likely from the same stream and indicating a site-specific dependence; and modeling reference condition with nutrient and land use data from all sites in the region may help better evaluate minimally disturbed conditions. 2) It does not seem likely that there are no instream or downstream effects of elevated nutrients because: phytoplankton blooms can occur during storm-free periods when waters slow and clear ; and downstream effects seem likely because patches of this ecoregion are so small and waters having to flow somewhere. 3) Criteria should be established to prevent dumping in this region, prevent degradation, and prevent surprises.

4. **MDEQ's criteria approach includes a Chl-a value of 125 mg/m^2 to be used as part of the related assessment information. Please comment on the selection of chlorophyll as the**

primary response variable, the derivation of the chlorophyll threshold, and its application as a statewide assessment indicator.

Using chlorophyll a or any indicator of valued ecological attributes, such as the diatom decreasers and Hilsenhoff biotic index, is an important check an assessments based on stressors because they directly address whether uses are being met. Chlorophyll a is a particularly important variable to use in determination of nutrient criteria and assessments of site compliance because it is probably the best indicator of algal biomass that we have and most effects of nutrients on designated uses of rivers and streams are caused by stimulation of either benthic or planktonic algal growth. MDEQ's derivation of 125 mg chl a m^{-2} as a management target to protect recreational use of rivers and aquatic life from DO stress is a model for what should be done by other states and tribes. In general, the chlorophyll standard was appropriately varied from region to region when reference condition nutrients were in the 30 µg TP/L and 300 µg TN/L range, but I do have concerns about using chlorophyll as an endpoint in the Rockies ecoregions where reference nutrient concentrations are low and nuisance levels of chlorophyll causing impairment of aesthetics and DO are not the only likely cause of changes in biological condition. Protecting biological condition at near natural levels may, however, be above the level of protection that stakeholders support in Montana. Although, tiered uses or an outstanding resource waters protection could be used to protect at least some low nutrient systems from increased productivity and resulting changes in biological condition. Other than these overall comments, details supporting the comment for question 4 are covered under question 1.

5. **Section 4.0 outlines a process for determining reach-specific nutrient criteria. Please comment on MDEQ's proposed approach for deriving reach-specific values.**

There are special situations when establishing nutrient criteria based on regional reference condition may be too high or too low. In the case of the Georgetown Lake Dam, the state statues call for a recalibration because they won't alter the location of the intact. The flow weighted approach in Bozeman Creek, Hyalite Creek, and East Gallatin also seems sound. I do question the relaxation of TN criteria above the very low 100 µg/L reference condition to around 250 µg TN/L, again for protecting high quality waters in Bozeman Creek, Hyalite Creek, and East Gallatin. This is the same issue as discussed for very low TP conditions in many of the Rockies ecoregions, but protecting high levels of biological conditions is a different issue than whether reach-specific criteria were determined appropriately based on management endpoints related to algal biomass, DO stress, and aesthetics.

3. **MDEQ is proposing to allow TN and TP criteria to be exceeded 20% of the time and be considered supporting aquatic life uses. This frequency was derived based on analysis of the Clark Fork River chl-a data. Please comment on the proposed exceedance frequency and whether allowing the stated magnitudes to be exceeded 20% of the time would not result in adverse effects on aquatic life. This information is discussed in the State's Assessment Methodology.**

6. **Montana is proposing to interpret the numeric criteria using the Students t-test and binomial test to determine whether a stream segment is impaired. Please comment on the State's rationale for this approach.**

I want to address questions 3 and 6 together. I think they are related. They are kind-of statistical issues.

First, I'd expect that 20% or more of observed TP and TN conditions would exceed criteria levels at sites maintaining an average target condition of, for example 150 mg chl a m^{-2}. The way that the nutrient criteria have been developed is based on the relationships between nutrient concentrations and chlorophyll observed at a site (Dodds et al. 1997, 2002, 2006). I'm going to use one of my Figures from Stevenson et al. (2006, redrawn and rescaled) to illustrate this because the non-linear relationships illustrated in the Dodds et al. (1997) paper are too complex to illustrate these principles and the plotted data looks a bit off in Dodds et al. (2002), which may be related to the erratum of Dodds et al. (2006).

In the statistical models that we use, there is both variation in the measurements of the independent and dependent variables (see Figure 1, blue and red lines respectively could be standard error bars of predicted and measured values). For example, when the average nutrient concentration is 30 µg TP/L in Michigan streams, chl a is approximately 20 mg/m^2. Michigan streams are grazer dominated and little periphyton accumulates with increases in nutrients, unless *Cladophora* can escape grazer control. Note values in Michigan are much lower than Kentucky by almost an order of magnitude. MDEQ based nutrient criteria on algal-nutrient model predictions to maintain biomass at a specific level or lower – usually 125 mg chl a m^{-2} target. So we should expect that nutrient concentrations will sometimes exceed the 125 mg chl a m^{-2} concentrations of the model, because there is variation around the predicted value. Actually, I'd expect the exceedance frequency to approach 50% as observed average nutrient concentrations at a site approach the criterion.

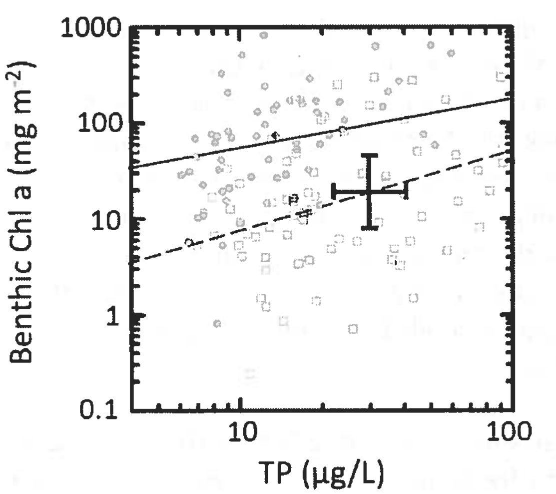

Figure 1. Figure 5 from Stevenson et al. (2006) redrawn and rescale. Open squares are values for streams in Michigan and shaded circles represent streams in Kentucky. Lines, dashed and solid respectively, are the linear relationships among the points.

This increase in exceedance frequency with nutrient concentration and algal biomass is illustrated in Suplee et al. (2011) Figure A4-1. Even for sites with the three lowest nutrient exceedance frequencies (and three lowest nutrient concentrations), maximum summer chl a is frequently greater than the 150 mg/m^2 expectation for the Clark Fork. The exceedance frequency actually provides a measure of risk of losing an attribute, in this case it's an aesthetically pleasing recreational venue and the potential for a DO event, which depending upon severity and extent, could have long-term repercussions for some biota. MDEQ have provided a margin of safety with lower biomass targets than impair aesthetics or cause DO stress and often lower nutrient creiteria than predicted to generate the target chlorophyll concentrations. This margin of safety may be the reason for average exceedance frequency at sites meeting uses being less than 50% even with nutrients concentrations near criteria.

Finally, I address the t-test and binomial test issues and issues with frequency distributions based on observations versus means. These issues are related to the risk of use support being affected by the way we use statistics to define reference condition and determine whether a site complies with water quality criteria. I want to start this discussion by reviewing a rationale for using frequency distributions of observations from reference conditions and a mean from a test sites to assess compliance at the test site. Then I'll transfer those concepts to evaluate how MDEQ's approach affects risk of supporting designated uses based on using observations, means, regression, binomial tests, and t-tests.

Consider the following scenario. Reference sites are selected because they are minimally disturbed based on land use and they support a specific level of aquatic life (and/or other uses) with an acceptable risk (let's say a management endpoint like chlorophyll a exceeds criteria 10% of the time). A frequency distribution is used to characterize central tendency and variation in a stressor (e.g. nutrients) that affect designated use within the range of conditions at reference sites. The frequency distribution is based on single samples from reference sites within an ecoregion (single independent observations in a statistical sense, Y_i, Figure 2A). We then use the 75th percentile of the frequency distribution of observed stressor conditions at references sites (Y_i^{75}) as a criterion, because we recognize that conditions vary around the average or median condition for reference sites due to spatial and temporal variation related to weather, flow, time of day, etc. and measurement error. Conceptually we use the 75th percentile of the frequency distribution of observed stressor conditions at references sites (Y_i^{75}) as a criterion because we feel that it's protective. Why? Well, one reason may be the statistical rule related to hypothesis testing called the 75% error bound. This is a rule of thumb that you can use to compare two means such that: if the mean of one sample (i.e. group) of observations (test sites) is outside the 75th confidence interval of the other sample of observations (reference sites), you can assume that there is little

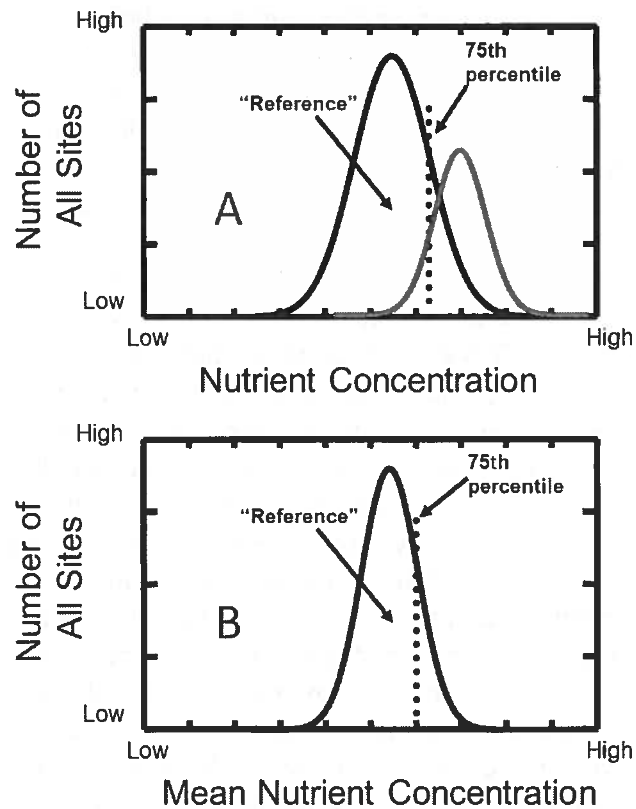

Figure 2. Distribution of nutrient concentrations when concentrations are represented by single samples site (A) and when concentrations are represented by means of samples at a site (B). The black distributions are reference site distributions. The blue distribution is the distribution of observations from a test site that is theoretically greater than the 75th percentile of the reference distribution according to a t-test or a binomial test.

probability that means of the first and second samples (groups of test and reference sites) are equal. Basically, 0.25 (1-0.75) is the attained significance for the difference between two means if the sample size is just 2 (if I remember correctly). So the idea is that conditions in the test set of sites would be different than the reference sites if the mean of test sites was greater than the 75th percentile of reference sites (based on single observations per site, or multiple observations from a smaller set of site that we could assume were independent). If agencies based development of criteria on mean measurements from a site, then the variance of the mean (Figure 2B) is much smaller than the variance of observations (Figure 2A). Comparing mean conditions at a site to the 75th percentile of a frequency distribution of mean observations at a site would be overprotective. Testing that mean of the test sites is significantly greater than the 75th percentile of reference sites (blue distribution in Figure 2A) versus just significantly greater than the mean of the reference site, would be underprotective.

MDEQ has proposed nutrient criteria that are the 75th percentile of reference condition or a higher concentration that is predicted to produce an effect that is undesireable. If my understanding of this process is correct, then using a t-test to determine whether mean conditions are greater than nutrient criteria would be underprotective, i.e. exceedance frequencies would be very high at sites before a site was found to be noncompliant. The mean concentration at the test site would have to be greater than the 75th percentile of the reference condition or the predicted level of nutrients causing a problem by an amount related to the variance in observed nutrient concentrations at the test site, the number of samples from the test site (n), and a t-statistic (which has a value of 2 when n is high). Issues associated with a binomial test are similar – some proportion of observed test site nutrient concentrations greater than 50% have to be greater than the criterion that is set at the maximum concentration that protects designated uses.

To counter these statistical issues with use of a t-stat causing underprotection of test sites, MDEQ does seem to have employed some margin of safety in setting the criteria at 125 mg chl a m^{-2}, which is below the 150 mg maximum okay level. In addition, MDEQ has adjusted acceptable levels of type I and II errors to reduce the problem of not detecting problems when they exist, and MDEQ is using chl a, a diatom indicator, and an HBI response criteria when nutrient concentrations are not obviously high. These additional rules can generate either greater under- or over-protection of waters. If only one the five criteria (N, P, chl, diatoms, HBI) has to fail, then that makes the assessment of compliance more protective than if the two nutrient criteria or all nutrient and biocriteria have to fail. I was not able to find details about compliance rules, which are apparently embedded in the spreadsheet that is referred to, so I can't evaluat Tier II assessments later. Tier II assessments, in which additional samples are collected and information is gathered also improves detection of non-compliance, reducing the error variation around means for the test site, which reduces possible difference between means at the test site and the nutrient criterion.

Overall, if I were a stakeholder concerned about protecting valued attributes within the streams, I'd be more concerned about potentially high risk of frequent loss of valued conditions (exceedance frequencies) when criteria are set at stressor-response model predictions that are too close to unacceptable levels of conditions, not the 20% exceedance problem or the t-test issue. The next frontier in deriving nutrient criteria may be bringing in a stronger risk assessment (e.g. Paul and McDonald 2005).

References

Cao, Y., C. P. Hawkins, J. Olson, and M. A. Kosterman. 2007. Modeling natural environmental gradients improves the accuracy and precision of diatom-based indicators. Journal of the North American Benthological **26**:566-585.

Davies, S. P., and S. K. Jackson. 2006. The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. Ecological Applications 16:1251-1266.

Dodds, W. K., V. H. Smith, and B. Zander. 1997. Developing nutrient targets to control benthic chlorophyll levels in streams: A case study of the Clark Fork River. Water Research 31:1738-1750.

Dodds, W. K., V. H. Smith, and K. Lohman. 2002. Nitrogen and phosphorus relationships to benthic algal biomass in temperate streams. Canadian Journal of Fisheries and Aquatic Sciences 59:865-874.

Dodds, W. K., and R. M. Oakes. 2004. A technique for establishing reference nutrient concentrations across watersheds affected by humans. Limnology and Oceanography: Methods 2:331-341.

Dodds, W. K., V. H. Smith, and K. Lohman. 2006. Erratum: Nitrogen and phosphorus relationships to benthic algal biomass in temperate streams. Canadian Journal of Fisheries and Aquatic Science 59:865-874.

Dodds, W. K., W. H. Clements, K. Gido, R. H. Hilderbrand, and R. S. King. 2010. Thresholds, breakpoints, and nonlinearity in freshwaters as related to management. Journal of the North American Benthological Society 29:988-997.

Hawkins, C. J., Y. Cao, and B. Rober. 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology **55**:1066-1085.

Herlihy A.T. and J.C. Sifneos. 2008. Developing nutrient criteria and classification schemes for wadeable streams in the conterminous US. Journal of the North American Benthological Society 27:932-948.

Muradian, R. 2001. Ecological thresholds: a survey. Ecological Economics 38:7-24.

Paul, J. F., and M. E. McDonald. 2006. Development of empirical, geographically specific water quality criteria: a confidential probability analysis approach. Journal of the American Water Resources Association 41:1211-1223.

Smith, A. J., R. W. Bode, and G. S. Kleppel. 2007. A nutrient biotic index (NBI) for use with benthic macroinvertebrate communities. Ecological Indicators **7**:371-386.

Soranno, P.A., K.S. Cheruvelil, R.J. Stevenson, S.L. Rollins, S.W. Holden, S. Heaton, and **E**. Torng. 2008. A framework for developing ecosystem-specific nutrient criteria: Integrating biological thresholds with predictive modeling. Limnology and Oceanography 53:773-787.

Stevenson, R.J. and S. Sabater. 2010. Understanding effects of global change on river ecosystems: science to support policy in a changing world. Hydrobiologia 657:3-18.

Stevenson, R. J., B. C. Bailey, M. C. Harass, C. P. Hawkins, J. Alba-Tercedor, C. Couch, S. Dyer, F. A. Fulk, J. M. Harrington, C. T. Hunsaker, and R. K. Johnson. 2004. Designing data collection for ecological assessments. In: M. T. Barbour, S. B. Norton, H. R. Preston, and K. W. Thornton, eds. Ecological Assessment of Aquatic Resources: Linking Science to

Decision-Making. Pgs 55-84. Society of Environmental Toxicology and Chemistry, Pensacola, Florida. ISBN 1-880611-56-2.

Stevenson, R. J., B. C. Bailey, M. C. Harass, C. P. Hawkins, J. Alba-Tercedor, C. Couch, S. Dyer, F. A. Fulk, J. M. Harrington, C. T. Hunsaker, and R. K. Johnson. 2004. Interpreting results of ecological assessments. In: M. T. Barbour, S. B. Norton, H. R. Preston, and K. W. Thornton, eds. Ecological Assessment of Aquatic Resources: Linking Science to Decision-Making. Pgs 85-111. Society of Environmental Toxicology and Chemistry, Pensacola, Florida. ISBN 1-880611-56-2.

Stevenson, R.J., S.T. Rier, C.M. Riseng, R.E. Schultz, and M.J. Wiley. 2006. Comparing effects of nutrients on algal biomass in streams in 2 regions with different disturbance regimes and with applications for developing nutrient criteria. Hydrobiologia 561:149-165.

Stevenson, R.J., B.E. Hill, A.T. Herlihy, L.L. Yuan, and S.B. Norton. 2008. Algal-P relationships, thresholds, and frequency distributions guide nutrient criterion development. Journal of the North American Benthological Society 27:783-799.

Stevenson, R. J., B. J. Bennett, D. N. Jordan, and R. D. French. 2012. Phosphorus regulates stream injury by filamentous algae, DO, and pH with thresholds in responses. Hydrobiologia 695:25-42.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. Ecological Applications 16:1267-1276.

Suplee, M.W. and V. Watson. 2012. Scientific and Technical Basis of the Numeric Nutrient Criteria for Montana's Wadeable Streams and Rivers—Addendum 1. Helena, MT: Montana Dept. of Environmental Quality.

Suplee, M.W., and R. Sada de Suplee. 2011 Assessment Methodology for Determining Wadeable Stream Impairment Due to Excess Nitrogen and Phosphorus Levels. Helena, MT: Montana Dept. of Environmental Quality.

Suplee, M.W., R. Sada de Suplee, D. Feldman and T. Laidlaw. 2005. Identification and Assessment of Montana Reference Streams: A Follow-up and Expansion of the 1992 Benchmark Biology Study. Helena, MT: Montana Dept. of Environmental Quality.

Yuan, L. L. 2004. Assigning macroinvertebrate tolerance classifications using generlised additive models. Freshwater Biology **49**:662-677.

Zar, J. H. 1974. Biostatistical Analysis. Prent